Minimax Structured Normal Means Inference

Akshay Krishnamurthy [*1]

[1]Microsoft Research, New York, NY 20011

January 27, 2016

Abstract

We provide a unified treatment of a broad class of noisy structure recovery problems, known as structured normal means problems. In this setting, the goal is to identify, from a finite collection of Gaussian distributions with different means, the distribution that produced some observed data. Recent work has studied several special cases including sparse vectors, biclusters, and graph-based structures. We establish nearly matching upper and lower bounds on the minimax probability of error for *any* structured normal means problem, and we derive an optimality certificate for the maximum likelihood estimator, which can be applied to many instantiations. We also consider an experimental design setting, where we generalize our minimax bounds and derive an algorithm for computing a design strategy with a certain optimality property. We show that our results give tight minimax bounds for many structure recovery problems and consider some consequences for interactive sampling.

1 Introduction

The prevalence of high-dimensional signals in modern scientific investigation has inspired an influx of research on recovering *structural information* from noisy data. These problems arise across a variety of scientific and engineering disciplines; for example identifying cluster structure in communication or social networks, multiple hypothesis testing in genomics, or anomaly detection in sensor networking. Specific structural assumptions include sparsity [13], low-rankedness [11], cluster structure [15], and many others [8].

The literature in this direction focuses on three inference goals: detection, localization or recovery, and estimation or denoising. Detection tasks involve deciding whether an observation contains some meaningful information or is simply ambient noise, while recovery and estimation tasks involve more precisely characterizing the information contained in a signal. These problems are closely related, but also exhibit important differences, and this paper focuses on the recovery problem, where the goal is to identify, from a finite collection of signals, which signal produced the observed data.

One frustration among researchers is that algorithmic and analytic techniques for these problems differ significantly for different structural assumptions. This issue was recently resolved in the context of the estimation, where the *atomic norm* [8] has provided a unifying algorithmic and analytical framework, but no such theory is available for detection and recovery problems. In this paper, we provide a unification for the recovery problem, leading to deeper understanding of how signal structure affects statistical performance.

---

[*]akshaykr@cs.cmu.edu

Modern measurement technology also often provides flexibility in designing strategies for data acquisition, and this adds an element of complexity to inference tasks. Data acquisition by both interactive and non-interactive experimental design is the typical situation in domains ranging from network tomography to crowdsourcing, but the statistical implications of these techniques are not fully understood. This paper also considers the experimental design setting and provides a generic solution and analysis of non-interactive experimental design for structure recovery problems.

To concretely describe our main contributions, we now develop the decision-theoretic framework of this paper. We study the **structured normal means problem** defined by a finite collection of vectors $\mathcal{V} = \{v_j\}_{j=1}^M \subset \mathbb{R}^d$ that index a family of probability distributions $\mathbb{P}_j = \mathcal{N}(v_j, I_d)$. An estimator T for the family $\mathcal{V}$ is a measurable function from $\mathbb{R}^d$ to $[M]$, and its maximum risk is:

$$\mathcal{R}(T, \mathcal{V}) = \sup_{j \in [M]} \mathcal{R}_j(T, \mathcal{V}), \qquad \mathcal{R}_j(T, \mathcal{V}) = \mathbb{P}_j[T(y) \neq j],$$

where we always use $y \sim \mathbb{P}_j$ to be the observation. We are interested in the **minimax risk**:

$$\mathcal{R}(\mathcal{V}) = \inf_T \mathcal{R}(T, \mathcal{V}) = \inf_T \sup_{j \in [M]} \mathbb{P}_j[T(y) \neq j]. \tag{1}$$

We call this the **isotropic setting** because each gaussian has spherical covariance. We are specifically interested in understanding how the family $\mathcal{V}$ influences the minimax risk. This setting encompasses recent work on sparsity recovery [13], biclustering [15, 7], and many graph-based problems [23]. An important example to keep in mind is the k-sets problem, where the collection $\mathcal{V}$ is formed by vectors $\mu \mathbf{1}_S$ for subsets $S \subset [d]$ of size k and some signal strength parameter μ. Instantiation of our results to this example will determine the critical scaling of μ in terms of d and k that is necessary and sufficient for achieving asymptotically zero minimax risk.

In the **experimental design setting**, the statistician specifies a sensing strategy, defined by a vector $B \in \mathbb{R}_+^d$. Using this strategy, under $\mathbb{P}_j$, the observation is, for each $i \in [d]$:

$$y(i) \sim v_j(i) + B(i)^{-1/2} \mathcal{N}(0, 1) = \mathcal{N}(v_j(i), B(i)^{-1}). \tag{2}$$

If $B(i) = 0$, then we say that $y(i) = 0$ almost surely. We call this distribution $\mathbb{P}_{j,B}$, to denote the dependence both on the target signal v_j and the sensing strategy B. The total measurement effort, or *budget*, used by the strategy is $\|B\|_1$, and we are interested in signal recovery under a budget constraint. Specifically, the minimax risk in this setting is:

$$\mathcal{R}(\mathcal{V}, \tau) = \inf_{T, B : \|B\|_1 \leq \tau} \sup_{j \in [M]} \mathbb{P}_{j,B}[T(y) \neq j]. \tag{3}$$

With this background, we now state our main contributions:

1. We give nearly matching upper and lower bounds on the minimax risk for both isotropic and experimental design settings (Theorems 2 and 5). This result matches many special cases that we are aware of [23]. Moreover, in examples with an asymptotic flavor, including the k-sets example, this shows that the maximum likelihood estimator (MLE) achieves the minimax rate.

2. In the isotropic case, we derive a condition on the family $\mathcal{V}$ under which the MLE exactly achieves the minimax risk, thereby certifying optimality of this estimator.

3. We give sufficient conditions that certify optimality of an experimental design strategy and also give an algorithm for computing such a strategy prior to data acquisition.

4. Lastly, we provide many examples to demonstrate the generality and applicability of our results.

We adopt the following notation: For a natural number M, we use $[M]$ to denote the set $\{1, \ldots, M\}$. For a sequence of problems indexed by a natural number $n \in \mathbb{N}$ and a signal strength parameter μ, we often state results in terms of the *minimax rate* and use the notation $\mu \asymp f(n)$ to denote this asymptotic scaling. This notation means that if $\mu = \omega(1)f(n)$, then the minimax risk can be driven to zero, while if $\mu = o(1)f(n)$, then the minimax risk approaches one asymptotically. Finally, for vectors $v, M \in \mathbb{R}^d$, we use $\|v\|_M = \sqrt{v^T \text{diag}(M)v}$ to denote the Mahalanobis norm.

2   Related Work

The structured normal means problem has a rich history in statistics and recent attention has focused on combinatorial structures. This line is motivated by statistical applications involving complex data sources, such as tasks in graph-structured signal processing [22]. Focusing on detection problems, a number of papers study various combinatorial structures, including k-sets [1], cliques [23], paths [4], and clusters [22] in graphs. While there are comprehensive results for many of these examples, a unifying theory for detection problems is still undeveloped.

Turning to recovery or localization, again several specific examples have been analyzed. The most popular example is the biclustering problem [15, 7, 23], which we study in Section 4.2. However, apart from this example and a few others [23], minimax bounds for the recovery problem are largely unknown. Moreover, we are unaware of a broadly applicable analysis like the method we develop here.

A unified treatment is possible for estimation problems, where the atomic norm framework gives sharp phase transitions on the mean squared error of the maximum likelihood estimator [8, 3, 20]. The atomic norm is a generic approach for encoding structural assumptions by decomposing the signal into a sparse convex combination of a set of base atoms (e.g., one-sparse vectors). While this line primarily focuses on linear inverse problems [8, 3], there are results for the estimation problem [20, 6], although neither set of results gives lower bounds on the minimax risk. Note that this approach is based on convex relaxation, and it is not immediate that such a relaxation will succeed for the recovery problem, as the probability of error for any dense family is one. Relatedly, the non-convexity of our risk poses new challenges that do not arise with the mean squared error objective.

The recovery problem we consider here has also been extensively studied in the signal processing and information theory literature, where it is referred to as Gaussian detection, or decoding with Additive White Gaussian Noise (AWGN), although the motivation and results are quite different. As the goal in channel coding is to reliably transmit as many bits of information as possible across a noisy channel, the vast majority of channel coding results focus on *codebook design* [19]. Researchers have studied structured (random and non-random) codebooks solely for computational efficiency, as random codes achieve optimal transmission rates but lead to a computationally intractable decoding problem. In contrast, in our setting the structured codebook is inherent to the problem and the main object of interest; the analyst has no control over the codebook and wants to achieve optimal decoding performance for the codebook specified.

Nevertheless, one line of work from this community analyzes the error probability for the maximum likelihood estimator/decoder for a given codebook (See [21] for a survey). Classical upper bounds include the min-distance bound and the Gallager bound [12], although the bound we prove here is also well-known [14]. Lower bounds come in two flavors: (a) sphere-packing lower bounds and (b) lower bounds on the maximum-likelihood error probability. The former is a lower bound that is independent of the particular codebook, so it does not give tight bounds on any specific family of vectors, while the latter applies only to the MLE, so it does not relate to the minimax risk. In contrast, our technique simultaneously applies to any codebook and

any estimator, leading to precise lower bounds on the minimax risk. To our knowledge, apart from the upper bound in Theorem 2, the results proved here do not appear in the information theory literature.

Turning briefly to the experimental design setting, a number of recent advances aim to quantify the statistical improvements enabled by experimental or interactive design in specific normal means instantiations [13, 23]. A unifying, interactive algorithm was recently proposed in the bandit optimization setting [9] but it is not known to improve on non-interactive approaches for the setting we consider. This work makes important progress, but a general-purpose interactive algorithm and a satisfactory characterization of the advantages offered by interactive sampling remain elusive open questions. This paper makes progress on the latter by developing lower bounds against all non-interactive approaches.

3 Main Results

In this section we develop the main results of the paper. We start by bounding the minimax risk in the isotropic setting, then develop a certificate of optimality for the maximum likelihood estimator. Lastly, we turn to the experimental design setting. We provide proofs in Appendix A.

3.1 Bounds on the Isotropic Minimax Risk

In the isotropic case, recall that we are given a finite collection $\mathcal{V}$ of vectors $\{v_j\}_{j=1}^M$ and an observation $y \sim \mathcal{N}(v_j, I_d)$ for some $j \in [M]$. Given such an observation, a natural estimator is the maximum likelihood estimator (MLE), which outputs the index j for which the observation was most likely to have come from. This estimator is defined as:

$$T_{\text{MLE}}(y) = \underset{j \in [M]}{\operatorname{argmax}} \, \mathbb{P}_j(y) = \underset{j \in [M]}{\operatorname{argmin}} \, \|v_j - y\|_2^2. \tag{4}$$

We will analyze this estimator, which partitions $\mathbb{R}^d$ based on a Voronoi Tessellation of the set $\mathcal{V}$.

As stated, the running time of the estimator is $O(Md)$, but it is worth pausing to remark briefly about computational considerations. In many examples of interest, the class $\mathcal{V}$ is combinatorial in nature, so M may be exponentially large, and efficient implementations of the MLE may not exist. However, as our setup does not preclude unstructured problems, the input to the estimator is the complete collection $\mathcal{V}$, so the running time of the MLE is linear in the input size. If the particular problem is such that $\mathcal{V}$ can be compactly represented (e.g. it has combinatorial structure), then the estimator may not be polynomial-time computable. This presents a real issue, as researchers have shown that a minimax-optimal polynomial time estimator is unlikely to exist for the biclustering problem [10, 18], which we study in Section 4. However, since the primary interest of this work is statistical in nature, we will ignore computational considerations for most of our discussion.

Our first result is a generic characterization of the minimax risk, which involves analysis of the MLE. The following function, which we call the **Exponentiated Distance Function**, plays a fundamental role.

Definition 1. *For a family $\mathcal{V}$ and $\alpha > 0$, the **Exponentiated Distance Function (EDF)** is:*

$$W(\mathcal{V}, \alpha) = \max_{j \in [M]} W_j(\mathcal{V}, \alpha) \tag{5}$$

$$\textit{with} \qquad W_j(\mathcal{V}, \alpha) = \sum_{k \neq j} \exp\left(\frac{-\|v_j - v_k\|_2^2}{\alpha} \right) \tag{6}$$

4

In the following theorem, we show that the EDF governs the performance of T_{MLE}. More importantly, this function also leads to a lower bound on the minimax risk, and the combination of these two statements shows that the MLE is nearly optimal for *any* structured normal means problem.

Theorem 2. *Fix $\delta \in (0, 1)$. If $W(\mathcal{V}, 8) \leq \delta$, then $\mathcal{R}(\mathcal{V}) \leq \mathcal{R}(\mathcal{V}, T_{MLE}) \leq \delta$. On the other hand, if $W(\mathcal{V}, 2(1 - \delta)) \geq 2^{\frac{1}{1-\delta}} - 1$, then $\mathcal{R}(\mathcal{V}) \geq \delta$.*

By setting $\delta = 1/2$ above, the second statement in the theorem may be replaced by: If $W(\mathcal{V}, 1) \geq 3$, then $\mathcal{R}(\mathcal{V}) \geq 1/2$. This setting often aids interpretability of the lower bound. Notice that the value of α disagrees between the lower and upper bounds, and this leads to a gap between the necessary and sufficient conditions. This is not purely an artifact of our analysis, as there are many examples where the MLE does not exactly achieve the minimax risk.

However, most structured normal means problems of interest have an asymptotic flavor, specified by a sequence of problems $\mathcal{V}_1, \mathcal{V}_2, \ldots$, and a signal-strength parameter μ, with observation $y \sim \mathcal{N}(\mu v_j, I_d)$ for some signal v_j in the current family. In this asymptotic framework, we are interested in how μ scales with the sequence to drive the minimax risk to one or zero. Almost all existing examples in the literature are of this form [23], and in all such problems, Theorem 2 shows that the MLE achieves the minimax rate. To our knowledge, such a comprehensive characterization of recovery problems is entirely new.

Note that the quantity $\mathcal{R}(\mathcal{V}, T_{\text{MLE}})$ is simply the worst case probability of error for the MLE, which has been extensively studied in the information theory community. Classical upper bounds on this quantity include the min-distance bound and Gallager's bound [12], but the EDF-based bound here has also appeared in the literature [14]. It is well known that the min-distance bound is often extremely loose (see Section 4), while application of Gallager's bound often involves challenging calculations [21]. The main novelty in our result is the lower bound, which shows that the EDF also leads to a lower bound on the error probability for all estimators/decoders, generically for all families $\mathcal{V}$. This new lower bound, accompanied with the existing upper bound, establishes that the EDF is the fundamental quantity in characterizing the minimax risk in these recovery problems.

Application of Theorem 2 to instantiations of the structured normal means problem requires bounding the EDF, which is significantly simpler than the typical derivation of this style of result. In particular, proving a lower bound no longer requires construction of a specialized subfamily of $\mathcal{V}$ as was the de facto standard in this line of work [15, 23]. In Section 4, we show how simple calculations can recover existing results. We also show how the EDF-based bound often gives much sharper results than the min-distance bound.

Turning to the proof, the EDF arises naturally as an upper bound on the failure probability of the MLE after applying a union bound and a Gaussian tail bound. Obtaining a lower bound based on the EDF is more challenging, and our proof is based on application of Fano's Inequality. We use a version of Fano's Inequality that allows a non-uniform prior and explicitly construct this prior using the EDF. This leads to our more general lower bound.

3.2 Minimax-Optimal Recovery

Theorem 2 shows that that maximum likelihood estimator achieves rate-optimal performance for *all* structured normal means recovery problems with the asymptotic flavor described. However, in some cases the MLE does not achieve the exact minimax risk, and it is therefore not the optimal estimator. In this section, we derive a sufficient condition for the *exact* minimax optimality of the MLE. As we will see via examples, the MLE is minimax optimal for several well-studied instantiations of the structured normal means problem, although analytically calculating the minimax risk is challenging.

The sufficient condition for optimality depends on a particular geometric structure of the family $\mathcal{V}$:

Definition 3. *A family $\mathcal{V}$ is **unitarily invariant** if there exists a set of orthogonal matrices $\{R_i\}_{i=1}^N$ such that for each vector $v \in \mathcal{V}$, the set $\{R_i v\}_{i=1}^N$ is exactly $\mathcal{V}$.*

In other words, the instance $\mathcal{V}$ can be generated by applying the orthogonal transforms to any fixed vector in the collection. Unitarily invariant problems exhibit high degrees of symmetry, and our next result shows that this symmetry suffices to certify optimality of the MLE.

Theorem 4. *If $\mathcal{V}$ is unitarily invariant, then the MLE is minimax optimal.*

Some remarks about the theorem are in order:

1. This theorem reduces the question of optimality to a purely geometric characterization of $\mathcal{V}$ and, as we will see, many important problems are unitarily invariant. One common family of orthogonal matrices is the set of all permutation matrices on $\mathbb{R}^d$.
2. This result does not characterize the risk of the MLE; it only shows that no other estimator has better risk. Specifically, it does not provide an analytic bound that is sharper than Theorem 2. From an applied perspective, an optimality certificate for an estimator is more important than a bound on the risk as it helps govern practical decisions, although risk bounds enable theoretical comparison.
3. Lastly, the result is not asymptotic in nature but rather shows that the MLE achieves the *exact* minimax risk for a fixed family $\mathcal{V}$. We are not aware of any other results in the literature that certify optimality of the MLE under our measure of risk.

The proof of this theorem is based on showing that the point-wise risk $\mathcal{R}_j(T, \mathcal{V})$ for the MLE is constant across the hypotheses $j \in [M]$. This argument uses the unitary invariance property and an explicitly representation of the MLE as a collection of polyhedral acceptance regions, with one region per hypothesis. Finally, we employ a dual characterization of the minimax risk to show that if the point-wise risk functional is constant, then the MLE is minimax-optimal.

In problems where Theorem 4 can be applied, our results give a complete understanding of the isotropic case. We know that the MLE exactly achieves the minimax risk and Theorem 2 also gives upper and lower bounds that match asymptotically.

3.3   The Experimental Design Setting

We now turn to the experimental design setting, where the statistician specifies a strategy $B \in \mathbb{R}_+^d$ and receives observation $y \sim \mathbb{P}_{j,B}$ given by Equation 2. Our main insight is that the choice of B only changes the metric structure of $\mathbb{R}^d$, and this change can be incorporated into the proof of Theorem 2. Specifically, the likelihood for hypothesis j, under sampling strategy B is:

$$\mathbb{P}_j(y|B) = \prod_{i=1}^d \sqrt{\frac{B(i)}{2\pi}} \exp\left\{ \frac{-B(i)(v_j(i) - y(i))^2}{2} \right\}$$

and the maximum likelihood estimator is:

$$T_{\mathrm{MLE}}(y, B) = \underset{j \in [M]}{\operatorname{argmin}} \|v_j - y\|_B^2.$$

We port Theorem 2 to this setting and show the following:

Theorem 5. *Fix $\delta \in (0,1)$ and any sampling strategy B with $\|B\|_1 \leq \tau$. Define the **Sampling Exponentiated Distance Function** SEDF:*

$$W(\mathcal{V}, \alpha, B) = \max_{j \in [M]} \sum_{k \neq j} \exp\left(\frac{-\|v_j - v_k\|_B^2}{\alpha} \right) \tag{7}$$

If $W(\mathcal{V}, 8, B) \leq \delta$ then $\mathcal{R}(\mathcal{V}, \tau) \leq \mathcal{R}(\mathcal{V}, T_{MLE}(y, B)) \leq \delta$. Conversely, if $W(\mathcal{V}, 2(1-\delta), B) \geq 2^{\frac{1}{1-\delta}} - 1$, then $\inf_T \sup_{j \in [M]} \mathbb{P}_{j,B}[T(y) \neq j] \geq \delta$.

The structure of the theorem is almost identical to that of Theorem 2, but it is worth making some important observations. First, the theorem holds for any non-interactive sampling strategy $B \in \mathbb{R}_+^d$, so the upper bound is strictly more general than Theorem 2. Therefore, this theorem also gives bounds for the non-isotropic or heteroskedastic case with known, shared covariance. Secondly, any strategy can be used to derive an upper bound on the minimax risk, but the same is not true for the lower bound. Instead the lower bound is dependent on the strategy, so one must minimize over strategies to lower bound $\mathcal{R}(\mathcal{V}, \tau)$. Fortunately, since the SEDF is convex in B and the budget constraint is polyhedral, computing the best strategy can be solved by convex programming. This gives a new algorithm for designing sampling procedures.

Specifically, for any α, to compute an associated design strategy, we solve the convex program,

$$\text{minimize}_{B \in \mathbb{R}_+^d, \|B\|_1 \leq \tau} \max_{j \in [M]} \sum_{k \neq j} \exp\left(\frac{-\|v_j - v_k\|_B^2}{\alpha} \right), \tag{8}$$

to obtain the sampling strategy $\hat{B}$ that minimizes the SEDF. For example, solving Program 8 with $\alpha = 1$ results in a strategy $\hat{B}$, and if $W(\mathcal{V}, 1, \hat{B}) \geq 3$, then we know that the minimax risk $\mathcal{R}(\mathcal{V}, \tau)$ over all strategies is at least $1/2$. On the other hand, solving with $\alpha = 8$ to obtain a (different) sampling strategy $\hat{B}$ and then using $\hat{B}$ with the MLE would give the tightest upper bound on the risk attainable by our proof technique. In Section 4, we demonstrate an example where this optimization leads to a non-uniform sampling strategy that outperforms uniform sampling.

In general it is challenging to analytically certify that an allocation strategy $\hat{B}$ minimizes the SEDF, but in some cases it is possible. Since the SEDF is convex in B, specializing the first-order optimality conditions for the resulting convex program gives the following:

Proposition 6. *Let $\hat{B}$ be a sampling strategy with $\|\hat{B}\|_1 = \tau$, $S(\hat{B}) \subset \mathcal{V}$ be the set of hypotheses achieving the maximum in $W(\mathcal{V}, \alpha, \hat{B})$, and π be a distribution on $S(\hat{B})$. If the quantity,*

$$\mathbb{E}_{j \sim \pi} \sum_{k \neq j} (v_k(i) - v_j(i))^2 \exp(-\|v_k - v_j\|_{\hat{B}}^2),$$

is constant across $i \in [d]$, then $\hat{B}$ is a minimizer of $W(\mathcal{V}, \alpha, B)$ subject to $\|B\|_1 \leq \tau$.

In many cases, this result leads to analytic lower bounds. Specifically, the result is especially useful when $\hat{B}$ is uniform across the coordinates, and $S(\hat{B}) = [M]$, so that all of the hypotheses achieve the maximum. In this case, it often suffices to choose π to be uniform over the hypotheses and exploit the high degree of symmetry to demonstrate the condition holds. As we will see in Section 4, many examples studied in the literature exhibit the requisite symmetry for this proposition to be applied in a straightforward manner.

Note that Tanczos and Castro [23] establish a similar sufficient condition for the uniform sampling strategy to be optimal. Their result however is slightly less general in that it only certifies optimality for the

uniform sampling strategy, whereas ours, in principle, can be applied more universally. In addition, their result applies only to problems where the hypotheses are of the form $\mu \mathbf{1}_S$ for a collection of subsets while ours is more general. This generality is important for some examples in Section 4. The other main difference is that their approach is not based on the SEDF, so their result is not directly applicable here.

4 Examples

This section contains four instantiations of structured normal means problems, and concrete results easily attainable from our approach. These examples have the asymptotic flavor described before, where we are interested in how a signal strength parameter μ scales with a sequence of instances.

The first example, the k-sets problem, is well studied, and as a warmup, we show how our technique recovers existing results. The second and third examples are motivated by biclustering and hierarchical clustering applications; in both problems our techniques establish a lower bound against all non-interactive approaches, demonstrating separation between interactive and non-interactive procedures. In the hierarchical clustering case, this separation is new. Finally the fourth example is a graph-structured signal processing problem where we empirically show that uniform sampling can be sub-optimal. The requisite calculations for these examples are deferred to Appendix A.

4.1 k-sets

In the k-sets problem, we have $M = \binom{d}{k}$ and each vector $v_j = \mathbf{1}_{S_j}$ where $S_j \subset [d]$ and $|S_j| = k$. The observation is $y \sim \mathcal{N}(\mu v_j, I_d)$ for some hypothesis j.

Corollary 7. *The minimax rate for k-sets is $\mu \asymp \sqrt{\log(k(d-k))}$ and $\mu \asymp \sqrt{\frac{d}{\tau} \log(k(d-k))}$, with budget τ. In the isotropic case, the MLE is minimax optimal.*

This corollary follows simply by bounding the EDF for the k-sets problem using binomial approximations. Using Proposition 6, it is easy to verify that uniform sampling is optimal here, which immediately gives the second claim. Finally using the set of all permutation matrices and exploiting symmetry, we can easily verify that this class is unitarily invariant. These bound agrees with established results in the literature [23]. To contrast, the min-distance bound from classical coding theory would reveal that the MLE succeeds when $\mu = \omega(\sqrt{k \log(d/k)})$. This bound is polynomially worse than the one attained by our more-refined EDF-based bound.

4.2 Biclusters

The biclustering problem operates over $\mathbb{R}^{d \times d}$ with $M = \binom{d}{k}^2$. We parametrize the class $\mathcal{V}$ with two indices so that $v_{ij} = \mathbf{1}_{S_i} \mathbf{1}_{S_j}^T \in \{0,1\}^{d \times d}$ with $|S_i| = |S_j| = k$. The observation is $y \sim \mathcal{N}(\mu \text{vec}(v_{ij}), I_{d^2})$ for a hypothesis (i, j).

Corollary 8. *The minimax rate for biclusters is $\mu \asymp \sqrt{\frac{\log(k(d-k))}{k}}$ and $\mu \asymp \sqrt{\frac{d^2}{\tau k} \log(k(d-k))}$, with budget constraint τ. In the isotropic case, the MLE is minimax optimal.*

Our bounds agree with existing analyses of this class [15, 7, 23]. The biclustering problem is interesting because there is a simple interactive algorithm that succeeds if $\mu = \omega\left(\sqrt{\left(\frac{d^2}{\tau k^2} + \frac{d}{\tau}\right) \log d}\right)$, which is a

factor of $\sqrt{k}$ smaller than the lower bound established here, demonstrating concrete statistical gains from interactivity [23]. We provide an analysis of this interactive algorithm in Appendix A. Note also that our EDF-based bound is polynomially better than classical min-distance bound.

4.3   Hierarchical Clustering

We study a model for similarity-based hierarchical clustering from Balakrishnan et al. [5]. This model is known as the Constant Block Model (CBM) and the balanced version is parameterized by a number of objects n, a minimum cluster size m, both of which are powers of 2, and a separation parameter μ. The hierarchical clustering is a perfectly balanced binary hierarchy on n objects with minimum cluster size m, where within cluster similarities are exactly μ larger than the between-cluster similarities at each level of the hierarchy. The induced $n \times n$ similarity matrix is related to an *ultrametric* (We provide a formal definition in Appendix A).

Balakrishnan et al. [5] analyze a recursive spectral clustering algorithm on this model in the presence of Gaussian noise. By associating the similarity matrix of each possible hierarchical clustering with an element of $\mathcal{V}$, their setting is a special case of the structured normal means problem, and our results can be used to characterize the minimax rate.

Corollary 9. *In the CBM, if $\mu = o\left(\sqrt{\frac{\log(nm)}{m}}\right)$ and under budget constraint τ, if $\mu = o\left(\sqrt{\frac{n^2 \log(nm)}{m\tau}}\right)$, then the minimax risk remains bounded away from 0. In the isotropic case, the MLE is minimax optimal.*

This corollary is a strict generalization of the minimax analysis of Balakrishnan et al. [5] who only consider the case $m = n/2$. Moreover, the results of Tanczos and Castro [23] do not apply here as the family does not correspond to indicator vectors of some set system. Thus, our more general treatment enables analysis of this important settings.

Note however that we only prove a lower bound on the minimax risk here. This lower bound, combined with existing analysis, establishes *exponential* separation between interactive and non-interactive approaches for this hierarchical clustering setting. The interactive algorithm of Krishnamurthy et al. [17] can recover clusters of size $m = \Omega(\log^2 n)$ with $\tau = \Theta(n\text{polylog}(n))$ and $\mu = \Theta(1)$. On the other hand, Corollary 9 shows that, with these settings of μ and τ, no non-interactive algorithm can recovery clusters of size $m = o(n)$ which is exponentially worse than the interactive algorithm.

4.4   Stars

Let $G = (V, E)$ be a graph and let the edges be numbered $1, \ldots, d$. The class $\mathcal{V}$ is the set of all *stars* in the graph, that is the vector $v_j \in \{0, 1\}^d$ is the indicator vector of all edges emanating from the jth node in the graph. Again the observation is $y \sim \mathcal{N}(\mu v_j, I_d)$ for some $j \in [|V|]$.

Corollary 10. *In the stars problem if the ratio between the maximum and minimum degree is bounded by a constant, i.e. $\frac{deg_{\max}}{deg_{\min}} \leq c$, then the minimax rate is $\mu \asymp \sqrt{\frac{\log(|V| - deg_{\min})}{deg_{\min}}}$.*

Again this agrees with a recent result of Tanczos and Castro [23], who consider s-stars of the complete graph, formed by choosing a vertex, and then activating s of the edges emanating out of that vertex. The two bounds agree in the special case of the complete graph with $s = |V| - 1$, but otherwise are incomparable, as they consider different problem structures. Note that the degree requirement here is not fundamental in Theorem 2, but arises from problem-specific approximations.
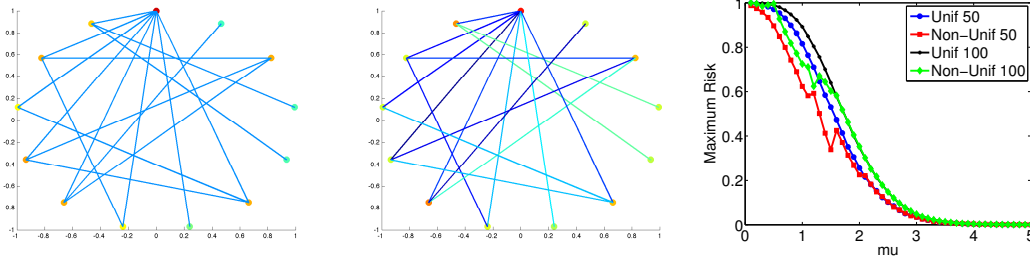
Figure 1: Left: A realization of the stars problem for a graph with 13 vertices and 34 edges with sampling budget $\tau = 34$. Edge color reflects allocation of sensing energy and vertex color reflects success probability for MLE under that hypothesis (warmer colors are higher for both). Isotropic (left) has minimum success probability of 0.44 and experimental design (center) has minimum success probability 0.56. Right: Maximum risk for isotropic and experimental design sampling as a function of μ for stars problem on a 50 and 100-vertex graph.

We highlight this example because the uniform allocation strategy does not necessarily minimize $W(\mathcal{V}, \alpha, B)$. In Figure 1, we construct a graph according to the Barabási-Albert model [2] and consider the class of stars on this graph. The simulation results show that optimizing the SEDF to find a sampling strategy is never worse than uniform sampling, and for low signal strengths it can lead to significantly lower maximum risk. Note that the risk for both uniform and non-uniform sampling approaches zero as $\mu \to \infty$, so for large μ, there is little advantage to optimizing the sampling scheme.

5   Discussion

This paper studies the structured normal means problem and gives a unified characterization of the minimax risk for isotropic and experimental design settings. Our work provides insights into how to choose estimators and how to design sampling strategies for these problems. Our lower bounds are critical in separating non-interactive and interactive sampling, which is an important research direction.

There are a number of exciting directions for future work, including extensions to other structure discovery problems, and to other observation models, such as compressive observations. We are most interested in developing a unifying theory for interactive sampling, analogous to the theory developed here. Another important and challenging direction is to consider recovery problems that involve nuisance parameters. We look forward to studying these problems in future work.

Acknowledgements

References

[1] Louigi Addario-Berry, Nicolas Broutin, Luc Devroye, and Gábor Lugosi. On combinatorial testing problems. *The Annals of Statistics*, 2010.

[2] Réka Albert and Albert-László Barabási. Statistical mechanics of complex networks. *Reviews of modern physics*, 2002.

[3] Dennis Amelunxen, Martin Lotz, Michael B. McCoy, and Joel A. Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference*, 2014.

[4] Ery Arias-Castro and Emmanuel J. Candès. Searching for a trail of evidence in a maze. *The Annals of Statistics*, 2008.

[5] Sivaraman Balakrishnan, Min Xu, Akshay Krishnamurthy, and Aarti Singh. Noise Thresholds for Spectral Clustering. In *Advances in Neural Information Processing Systems*, 2011.

[6] Badri Narayan Bhaskar, Gongguo Tang, and Benjamin Recht. Atomic norm denoising with applications to line spectral estimation. *IEEE Transactions on Signal Processing*, 2013.

[7] Cristina Butucea and Yuri I. Ingster. Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli*, 2013.

[8] Venkat Chandrasekaran, Benjamin Recht, Pablo A. Parrilo, and Alan S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 2012.

[9] Shouyuan Chen, Tian Lin, Irwin King, Michael R. Lyu, and Wei Chen. Combinatorial Pure Exploration of Multi-Armed Bandits. In *Advances in Neural Information Processing Systems*, 2014.

[10] Yudong Chen and Jiaming Xu. Statistical-computational tradeoffs in planted problems and submatrix localization with a growing number of clusters and submatrices. *arXiv:1402.1267*, 2014.

[11] David Donoho and Matan Gavish. Minimax risk of matrix denoising by singular value thresholding. *The Annals of Statistics*, 2014.

[12] Robert G Gallager. *Low-Density Parity-Check Codes*. PhD thesis, MIT, 1963.

[13] Jarvis Haupt, Rui Castro, and Robert Nowak. Distilled sensing: Adaptive sampling for sparse detection and estimation. *IEEE Transactions on Information Theory*, 2011.

[14] Hanan Herzberg and Gregory Poltyrev. Techniques of bounding the probability of decoding error for block coded modulation structures. *IEEE Transactions on Information Theory*, 1994.

[15] Mladen Kolar, Sivaraman Balakrishnan, Alessandro Rinaldo, and Aarti Singh. Minimax Localization of Structural Information in Large Noisy Matrices. *Advances in Neural Information Processing Systems*, 2011.

[16] Akshay Krishnamurthy. Minimaxity in Structured Normal Means Inference. *arXiv:1506.07902*, 2015.

[17] Akshay Krishnamurthy, Sivaraman Balakrishnan, Min Xu, and Aarti Singh. Efficient active algorithms for hierarchical clustering. In *International Conference on Machine Learning*, 2012.

[18] Zongming Ma and Yihong Wu. Computational barriers in minimax submatrix detection. *arXiv:1309.5914*, 2013.

[19] David JC MacKay. *Information theory, inference and learning algorithms*. Cambridge university press, 2003.

[20] Samet Oymak and Babak Hassibi. Sharp mse bounds for proximal denoising. *arXiv:1305.2714*, 2013.

[21] Igal Sason and Shlomo Shamai. *Performance analysis of linear codes under maximum-likelihood decoding: A tutorial*. Now Publishers Inc., 2006.

[22] James Sharpnack, Akshay Krishnamurthy, and Aarti Singh. Near-optimal Anomaly Detection in Graphs using Lovasz Extended Scan Statistic. *Advances in Neural Information Processing Systems*, 2013.

[23] Ervin Tánczos and Rui Castro. Adaptive sensing for estimation of structured sparse signals. *arXiv:1311.7118*, 2013.

A Proofs

A.1 Proof of Theorem 2

Analysis of MLE: We first analyze the maximum likelihood estimator:

$$T_{MLE}(y) = \operatorname*{argmin}_{j \in [M]} \|v_j - y\|_2^2$$

This estimator succeeds as long as $\|v_k - y\|_2^2 > \|v_{j^\star} - y\|_2^2$ for each $k \neq j^\star$, when $y \sim \mathbb{P}_{j^\star}$. This condition is equivalent to:

$$\|v_k - y\|_2^2 > \|v_{j^\star} - y\|_2^2 \Leftrightarrow \langle \epsilon, v_k - v_{j^\star} \rangle < \frac{1}{2}\|v_{j^\star} - v_k\|_2^2$$

where $\epsilon \sim \mathcal{N}(0, 1)$. This follows from writing $y = v_{j^\star} + \epsilon$ and then expanding the squares. So we must simultaneously control all of these events, for fixed $j^\star$:

$$\begin{aligned}
&\mathbb{P}_{\epsilon \sim \mathcal{N}(0, I_d)} \left[\forall k \neq j^\star. \langle \epsilon, v_k - v_{j^\star} \rangle < \|v_{j^\star} - v_k\|_2^2/2 \right] \\
&= 1 - \mathbb{P}_{\epsilon \sim \mathcal{N}(0, I_d)} \left[\exists k \neq j^\star. \langle \epsilon, v_k - v_{j^\star} \rangle \geq \|v_{j^\star} - v_k\|_2^2/2 \right] \\
&\geq 1 - \sum_{k \neq j^\star} \mathbb{P}_{\epsilon \sim \mathcal{N}(0, I_d)} \left[\langle \epsilon, v_k - v_{j^\star} \rangle \geq \|v_{j^\star} - v_k\|_2^2/2 \right]
\end{aligned}$$

By a gaussian tail bound, this probability is:

$$\mathbb{P}_{\epsilon \sim \mathcal{N}(0, I_d)} \left[\langle \epsilon, v_k - v_{j^\star} \rangle \geq \|v_{j^\star} - v_k\|_2^2/2 \right] \leq \exp \left\{ -\frac{1}{8} \|v_{j^\star} - v_k\|_2^2 \right\}$$

So that the total failure probability is upper bounded by:

$$\mathbb{P}_{j^\star}[\hat{j} = j^\star] \leq \sum_{k \neq j^\star} \exp \left\{ -\frac{1}{8} \|v_{j^\star} - v_k\|_2^2 \right\} = W_{j^\star}(\mathcal{V}, 8)$$

So if j is the truth, then the probability of error is smaller than δ when $W_j(\mathcal{V}, 8) \leq \delta$. For the maximal (over hypothesis choice j) probability of error to be smaller than δ, it suffices to have $W(\mathcal{V}, 8) \leq \delta$.

Fundamental Limit: We now turn to the fundamental limit. We start with a version of Fano's inequality with non-uniform prior.

Lemma 11 (Non-uniform Fano Inequality). *Let $\Theta = \{\theta\}$ be a parameter space that indexes a family of probability distributions P_θ over a space $\mathcal{X}$. Fix a prior distribution π, supported on Θ and consider $\theta \sim \pi$ and $X \sim P_\theta$. Let $f : \mathcal{X} \to \Theta$ be any possibly randomized mapping, and let $p_e = \mathbb{P}_{\theta \sim \pi, X \sim P_\theta}[f(X) \neq \theta]$ denote the probability of error. Then:*

$$p_e \geq 1 - \frac{\sum_\theta \pi(\theta) KL(P_\theta \| P_\pi) + \log 2}{H(\pi)}$$

where $P_\pi(\cdot) = \mathbb{E}_{\theta \sim \pi} P_\theta(\cdot)$ is the mixture distribution. In particular, we have:

$$\inf_f \sup_\theta \mathbb{P}_{X \sim P_\theta}[f(X) \neq \theta] \geq \inf_f \mathbb{E}_{\theta \sim \pi} \mathbb{P}_{X \sim P_\theta}[f(X) \neq \theta] \geq 1 - \frac{\sum_\theta \pi(\theta) KL(P_\theta \| P_\pi) + \log 2}{H(\pi)}$$

13

Proof. Consider the Markov Chain $\theta \to X \to \hat{\theta} \triangleq f(X)$ where $\theta \sim \pi$ and $X|\theta \sim P_\theta$. Let $E = \mathbf{1}[\hat{\theta} \neq \theta]$.

$$H(E|X) + H(\theta|E, X) = H(E, \theta|X) = H(\theta|X) + H(E|\theta, X) \geq H(\theta|X)$$

Now, $H(E|\theta, X) \geq 0$ and since conditioning only reduces entropy, we have the inequality

$$H(\theta|X) \leq H(p_e) + H(\theta|E, X) = H(p_e) + H(\theta|E = 0, X)P[E = 0] + H(\theta|E = 1, X)P[E = 1]$$
$$= H(p_e) + p_e H(\theta)$$

which proves the usual version of Fano's inequality. We want to write $H(\theta|X)$ in terms of the KL divergence, using the mixture distribution P_π.

$$H(\theta|X) = H(\theta, X) - H(X) = \int \sum_\theta \pi(\theta) P_\theta(x) \log \left(\frac{\sum_\theta \pi(\theta) P_\theta(x))}{\pi(\theta) P_\theta(x)} \right) dx$$
$$= \sum_\theta \pi(\theta) \int P_\theta(x) \log \left(\frac{P_\pi(x)}{P_\theta(x)} \right) dx - \sum_\theta \pi(\theta) \log \pi(\theta)$$
$$= -\sum_\theta \pi(\theta) KL(P_\theta || P_\pi) + H(\pi)$$

Combining these gives the bound:

$$H(p_e) + p_e H(\pi) \geq H(\pi) - \sum_\theta \pi(\theta) KL(P_\theta || P_\pi),$$

By upper bounding $H(p_e) \leq \log 2$ and rearranging we prove the claim. $\qquad\square$

For a distribution $\pi \in \Delta_{M-1}$ over the hypothesis, let $P_\pi(\cdot) = \sum_k \pi_k P_k(\cdot)$ be the mixture distribution. Then Fano's inequality (Lemma 11) states that the minimax probability of error is lower bounded by:

$$\mathcal{R}(\mathcal{V}) = \inf_T \sup_j \mathbb{P}_j[T(y) \neq j] \geq \inf_T \mathbb{E}_{j \sim \pi} \mathbb{E}_{y \sim j} \mathbf{1}[T(y) \neq j]$$
$$\geq 1 - \frac{\mathbb{E}_{k \sim \pi} KL(P_k || P_\pi) + \log 2}{H(\pi)}.$$

Fix $\delta \in (0, 1)$ and let $j^\star = \operatorname{argmax}_{j \in [M]} W_j(2(1 - \delta))$. We will use a prior based on this quantity:

$$\pi_k \propto \exp \left(-\frac{\|v_{j^\star} - v_k\|_2^2}{2(1 - \delta)} \right)$$

With this prior, the entropy becomes:

$$H(\pi) = \sum_k \pi_k \log \left(\frac{\sum_i \exp \left(-\frac{\|v_{j^\star} - v_i\|_2^2}{2(1-\delta)} \right)}{\exp \left(-\frac{\|v_{j^\star} - v_k\|_2^2}{2(1-\delta)} \right)} \right)$$
$$= \log(W(\mathcal{V}, 2(1 - \delta)) + 1) + \sum_k \pi_k \frac{\|v_{j^\star} - v_k\|_2^2}{2(1 - \delta)}$$
$$= \log(W(\mathcal{V}, 2(1 - \delta)) + 1) + \frac{1}{1 - \delta} \sum_k \pi_k KL(P_k || P_{j^\star})$$

The 1 inside the first $\log$ comes from the fact that in the definition $W_{j^\star}$, we do not include the term involving $j^\star$ in the sum, while our prior π does place mass proportional to 1 on hypothesis $j^\star$. The term involving the KL-divergence follows from the fact that the KL between two gaussians is one-half the ℓ_2^2-distance between their means.

Looking at the lower bound from Fano's inequality, we see that if:

$$\mathbb{E}_{k \sim \pi} KL(P_k || P_\pi) + \log 2 \leq (1 - \delta) H(\pi) = (1 - \delta) \log(W(\mathcal{V}, 2(1 - \delta)) + 1) + \sum_k \pi_k KL(P_k || P_{j^\star})$$

then the probability of error is lower bounded by δ. Of course it is immediate that:

$$\begin{aligned}
\sum_k \pi_k KL(P_k || P_{j^\star}) &= \sum_k \pi_k \int P_k(x) \log \left(\frac{P_k(x) P_\pi(x)}{P_\pi(x) P_{j^\star}(x)} \right) \\
&= \sum_k \pi_k \int P_k(x) \log \frac{P_k(x)}{P_\pi(x)} + \sum_k \int \pi_k P_k(x) \log \frac{P_\pi(x)}{P_{j^\star}(x)} \\
&= \sum_k \pi_k KL(P_k || P_\pi) + KL(P_\pi || P_{j^\star}) \geq \mathbb{E}_k KL(P_k || P_\pi)
\end{aligned}$$

So the condition reduces to requiring that:

$$\log 2 \leq (1 - \delta) \log(W(\mathcal{V}, 2(\delta - 1) + 1).$$

After some algebra, this is equivalent to:

$$W(\mathcal{V}, 2(\delta - 1)) \geq 2^{\frac{1}{1-\delta}} - 1 \qquad \square$$

A.2 Proof of Theorem 5

The proof of Theorem 5 is essentially the same as the proof of Theorem 2, coupled with two observations. First, for a sampling strategy $B \in \mathbb{R}_+^d$ the maximum likelihood estimator is:

$$T_{\mathrm{MLE}}(y, B) = \operatorname*{argmin}_{j \in [M]} \|v_j - y\|_B^2$$

so the analysis of the MLE depends on the Mahalanobis norm $\| \cdot \|_B$ instead of the ℓ_2 norm.

Similarly, the KL divergence between the distribution $\mathbb{P}_{j,B}$ and $\mathbb{P}_{k,B}$ depends on the Mahalanobis norm $\| \cdot \|_B$ instead of the ℓ_2 norm. Specifically, we have:

$$KL(\mathbb{P}_{j,B} || \mathbb{P}_{k,B}) = \frac{1}{2} \|v_j - v_k\|_B^2.$$

The lower bound proof instead use this metric structure, but the calculations are equivalent. $\qquad \square$

A.3 Proof of Proposition 6

To simplify the presentation, let $f(B) = W(\mathcal{V}, \alpha, B)$. $f(B)$ is convex and (strictly) monotonically decreasing, so we know that the minimum will be achieved when the constraint is tight, i.e. when $\|B\|_1 = \tau$. The Lagrangian is:

$$\mathcal{L}(B, \lambda) = f(B) + \lambda(\|B\|_1 - \tau)$$

and the minimum is achieved at $\hat{B}$, with $\|\hat{B}\|_1 = \tau$, if there is a value $\hat{\lambda}$ such that $0 \in \partial \mathcal{L}(\hat{B}, \hat{\lambda})$. Observing that the subgradient is $\partial f(B) + \lambda \mathbf{1}$, it suffices to ignore the Lagrangian term and instead ensure that $\partial f(B) \propto \mathbf{1}$. $f(B)$ is a maximum of M convex functions, where $f_j(B)$ is the function corresponding to hypothesis v_j, and, by direct calculation, the subgradient of this function $f_j(B)$ is:

$$\frac{\partial f_j(B)}{\partial B_i} = \sum_{k \neq j} -(v_k(i) - v_j(i))^2 \exp(-\|v_k - v_j\|_B^2).$$

Moreover, the subgradient of the maximum of a set of functions is the convex hull of the subgradients of all functions achieving the maximum. This means that if there exists a distribution π, supported over the maximizers of $f(\hat{B})$, such that the expectation of the subgradients is constant, we have certified optimality of $\hat{B}$. This is precisely the condition in the Proposition. □

A.4 Proof of Theorem 4

Our approach is based on a well-known connection between the minimax risk and the Bayes risk. For a structured normal means problem defined by a family $\mathcal{V}$, the *Bayes risk* for an estimator T under prior $\pi \in \Delta_{M-1}$ is given by:

$$B_\pi(T) = \sum_{j=1}^{M} \pi_j \mathbb{P}_j[T(y) \neq j].$$

We say that an estimator T is the Bayes estimator for prior π if it achieves the minimum Bayes risk. A simple calculation reveals the structure of the Bayes estimator for any prior π and this structural characterization is essential to our development.

Proposition 12. *For any prior π, the Bayes estimator T_π has polyhedral acceptance regions, that is the estimator is of the form:*

$$T(Y) = j \text{ if } y \in A_j,$$

with $A_j = \{x : \Gamma_j x \geq b_j\}$ and $\Gamma_j \in \mathbb{R}^{M \times d}$ has $v_j - v_k$ in the kth row and b_j has $\frac{1}{2}(\|v_j\|_2^2 - \|v_k\|_2^2) + \log \frac{\pi_k}{\pi_j}$ in the kth entry. These polyhedral sets A_j partition the space $\mathbb{R}^d$.

Proof. To prove Proposition 12, we make two claims. First we certify that for a prior π, the Maximum a Posteriori (MAP) estimator is a Bayes estimator for prior π. Given π, the map estimator is:

$$T_\pi(y) = \underset{j}{\arg\max}\, \pi(j) \exp\{-\|v_j - y\|_2^2/2\}$$

Define the posterior risk of an estimator T to be the expectation of the loss, under the posterior distribution on the hypothesis. In our case this is:

$$r(T|y) = \sum_{j=1}^{M} \mathbf{1}[T(y) \neq j]\pi(j|y) \qquad \text{where} \qquad \pi(j|y) \propto \pi(j) \exp\{-\|v_j - y\|_2^2/2\}.$$

For a fixed y, this quantity is minimized by letting $T(y)$ be the maximizer of the posterior, as this makes the $0 - 1$ loss term zero for the largest $\pi(j|y)$ value. Thus for each y we minimize the posterior risk by letting

16

$T(y)$ be the MAP estimate. The result follows by the well known fact that if an estimator minimizes the posterior risk at each point, then it is the Bayes estimator.

This argument shows that the only types of estimators we need to analyze are MAP estimators under various priors. This gives us the requisite structure to prove Proposition 12.

Specifically, for a prior π, for the MAP estimate to predict hypothesis j, it must be the case that:

$$\forall k \neq j. \qquad \pi_j \exp\{-\|v_j - y\|_2^2/2\} \geq \pi_k \exp\{-\|v_j - v_k\|_2^2/2\}.$$

This can be simplified to:

$$\langle v_j - v_k, y \rangle \geq \frac{1}{2}(\|v_j\|_2^2 - \|v_k\|_2^2) + \log \frac{\pi_k}{\pi_j}.$$

Thus the acceptance region for the hypothesis j is the set of all points y that satisfy all of these $M - 1$ inequalities. This is exactly the polyhedral set A_j. $\qquad \square$

We also exploit the relationship between the minimax risk and the Bayes risk. This is a well known result, where the prior π below is known as the *least-favorable prior*.

Proposition 13. *Suppose that T is a Bayes estimator for some prior π. If the risk $\mathcal{R}_j(T) = \mathcal{R}_{j'}(T)$ for all $j \neq j' \in [M]$, then T is a minimax optimal estimator.*

Proof. We provide a proof of this well-known result showing that the Bayes estimator with uniform risk landscape is minimax optimal. Let T_π be the Bayes estimator under prior π and let T_0 be some other estimator. Since T_π has constant risk landscape, we know that $\max_j \mathcal{R}_j(\mathcal{V}, T_\pi) = B_\pi(T_\pi)$, or the minimax risk for T_π is equal to its Bayes risk. We know that the Bayes risk of T_0 is at most the minimax risk for T_0, i.e. $B_\pi(T_0) \leq \max_j \mathcal{R}_j(\mathcal{V}, T_0)$. If it were the case that T_0 had strictly lower minimax risk, then we have:

$$B_\pi(T_0) \leq \max_j \mathcal{R}_j(\mathcal{V}, T_0) < \max_j \mathcal{R}_j(\mathcal{V}, T_\pi) \leq B_\pi(T_\pi).$$

However, this is a contradiction since T_π is the Bayes estimator under prior π, meaning that it minimizes the Bayes risk. $\qquad \square$

Equipped with these results, we now turn to the proof of the theorem.

Proof of Theorem 4: Our goal is to apply Proposition 13. By the fact that $\mathcal{R}_j(\mathcal{V}, T) = 1 - \mathbb{P}_j[A_j]$ where A_j is T's acceptance region for hypothesis j, we must show that the $\mathbb{P}_j$ probability content of the acceptance regions are constant. Ignoring the normalization factor of the gaussian density, this is:

$$\int_{A_j} \exp\{-\|v_j - x\|_2^2/2\}dx,$$

where $A_j = \{z | \Gamma_j z \geq b_j\}$ as defined in Proposition 12. We will exploit the unitary invariance of the family.

For any pair of hypothesis j, k, let R_{jk} be the orthogonal matrix such that $v_k = R_{jk}v_j$ and note that R_{kj}, the orthogonal matrix that maps v_k to v_j, is just R_{jk}^T. This also means that $R_{jk}R_{jk}^T = R_{jk}R_{kj} = I$. Via a change of variables $x = R_{kj}y$, the integrand becomes:

$$\exp\{-\|v_j - R_{kj}y\|_2^2/2\} = \exp\{-\|R_{jk}v_j - R_{jk}R_{kj}y\|_2^2/2\} = \exp\{-\|v_k - y\|_2^2/2\}.$$

Thus, we have translated to the P_k measure.

As for the region of integration, first note that since $v_i = R_{ji} v_j$, it must be the case that $\|v_j\|_2^2 = \|v_i\|_2^2$ for all $j, i \in [M]$. This means that the vector b_j defining the acceptance region, which for the MLE has coordinates $b_j(i) = \frac{1}{2}(\|v_j\|_2^2 - \|v_i\|_2^2)$, is just the all-zeros vector. The region of integration is therefore:

$$\{z|\Gamma_j z \geq 0\} = \{z|\Gamma_j R_{kj} z \geq 0\}.$$

We must check that this polytope is exactly A_k, which means that we must check that for each i, $(v_j - v_i)^T R_{kj}$ is a row of the Γ_k matrix. But:

$$(v_j - v_i)^T R_{kj} = v_j^T R_{jk}^T - v_i^T R_{jk}^T = v_k^T - v_i^T R_{jk}^T.$$

Since v_i can generate the family $\mathcal{V}$, it must be the case that $R_{jk} v_i \in \mathcal{V}$ so that this difference does correspond to some row of Γ_k. Since we apply the same unitary operator to all of the rows, it must be the case that the number of distinct rows is unchanged, or in other words, there is a bijection from the rows in $\Gamma_j R_{kj}$ to the rows in Γ_k. Therefore, the transformed region of integration, after the change of variable $x = R_{kj} y$ is exactly the acceptance region A_k, and the integrand is the $\mathbb{P}_k$ measure. This means that $\mathbb{P}_k[A_k] = \mathbb{P}_j[A_j]$ and this is true for all pairs (j, k), so that the risk landscape is constant. By Proposition 13, this certifies optimality of the MLE. $\qquad\square$

A.5 Calculations for the examples

Calculations for k-Sets: We must upper and lower bound $W(\mathcal{V}, \alpha)$. First note that by symmetry, every hypothesis achieves the maximum, so it suffices to compute just one of them:

$$W(\mathcal{V}, \alpha) = \sum_{k \neq j} \exp\left(-\|v_k - v_j\|_2^2/\alpha\right) = \sum_{s=1}^{k} \binom{k}{s} \binom{d-k}{s} \exp(-2s\mu^2/\alpha).$$

This follows by noting that the ℓ_2^2 distance between two hypothesis is the symmetric set difference between the two subsets, and then by a simple counting argument. Using well known bounds on binomial coefficients, we obtain:

$$W(\mathcal{V}, \alpha) \leq \sum_{s=1}^{k} \exp(s \log(ke/s) + s \log((d-k)e/s) - 2s\mu^2/\alpha)$$

$$= \sum_{s=1}^{k} \exp(s \log(e^2 k(d-k)/s^2) - 2s\mu^2/\alpha)$$

$$\leq k \exp(\log e^2 k(d-k) - 2\mu^2/\alpha) \qquad \text{if } 2\mu^2/\alpha \geq \log(e^2 k(d-k))$$

This is smaller than δ whenever $\mu^2 \geq \alpha \log(ek(d-k)/\delta)$, which subsumes the requirement above. For the lower bound:

$$W(\mathcal{V}, \alpha) \geq \sum_{s=1}^{k} \exp(s \log(k/s) + s \log((d-k)/s) - 2s\mu^2/\alpha) \geq \exp(-2\mu^2/\alpha + \log(k(d-k)))$$

which goes to infinity if $\mu^2 = o(\alpha \log(k(d-k)))$.

To certify that the uniform allocation strategy minimizes $W(\mathcal{V}, \alpha, B)$, we apply Proposition 6. Fix τ and let $\hat{B}$ be such that $\hat{B}(i) = \tau/d$. By symmetry, every hypothesis achieves the maximum under this allocation strategy, and we will take π to be the uniform distribution over all hypothesis.

For a hypothesis j and a coordinate i, the subgradient $\frac{\partial f_j(B)}{\partial B(i)}$ at $\hat{B}_i$ depends on the whether $v_j(i) = 0$ or not. If $v_j(i) = 0$, then:

$$\frac{\partial f_j(B)}{\partial B(i)} = \mu^2 \sum_{s=1}^{k} \binom{d-k-1}{s-1} \binom{k}{k-s} \exp(-2\tau\mu^2 s^2/d),$$

and if $v_j(i) = \mu^2$ then:

$$\frac{\partial f_j(B)}{\partial B(i)} = \mu^2 \sum_{s=1}^{k} \binom{d-k}{s} \binom{k-1}{k-s} \exp(-2\tau\mu^2 s^2/d).$$

Both of these follow from straightforward counting arguments. Notice that the value of the subgradient depends only on whether $v_j(i) = 0$ or not, and under the uniform distribution π, $\mathbb{E}_{j\sim\pi} v_j(i) = \mathbb{E}_{j\sim\pi} v_j(i')$. This implies that the constant vector is in the subgradient of $f(B)$ at $\hat{B}$, so that $\hat{B}$ is the minimizer of $W(\mathcal{V}, \alpha, B)$ subject to $\|B\|_1 \le \tau$.

We have already done the requisite calculation to bound the minimax risk under sampling. The calculations above show that if $\mu = \omega(\sqrt{\frac{d}{\tau} \log(k(d-k))})$ then the maximum likelihood estimator, when using the uniform sampling strategy has risk tending to zero. Conversely if $\mu = o(\sqrt{\frac{d}{\tau} \log(k(d-k))})$ then the minimax risk, for *any* allocation strategy tends to one.

Calculation for Biclusters: Due to symmetry, all hypotheses achieve the maximum and therefore, we can directly calculate $W(\mathcal{V}, \alpha)$. We use the notation C_n^i to denote the binomial coefficient $\binom{n}{i}$.

$$W(\mathcal{V}, \alpha) = \sum_{s_r=1}^{k} \sum_{s_c=1}^{k} C_k^{s_r} C_k^{s_c} C_{d-k}^{s_r} C_{d-k}^{s_c} \exp\left(-\frac{2\mu^2}{\alpha}(s_r(k-s_c) + s_c(k-s_r) + s_r s_c)\right)$$
$$+ \sum_{s_r=1}^{k} C_k^{s_r} C_{d-k}^{s_r} \exp\left(-\frac{2\mu^2}{\alpha}(s_r k)\right) + \sum_{s_c=1}^{k} C_k^{s_c} C_{d-k}^{s_c} \exp\left(-\frac{2\mu^2}{\alpha}(s_c k)\right)$$

This last two term comes from the case where $s_c = 0$ or $s_r = 0$, which is all of the hypotheses that share the same columns but disagree on the rows (or share the same rows but disagree on the columns). Using binomial approximations, the first term can be upper bounded by:

$$\le \sum_{s_r=1}^{k} \sum_{s_c=1}^{k} \exp\left(s_r \log\frac{k(d-k)e^2}{s_r^2} + s_c \log\frac{k(d-k)e^2}{s_c^2} - \frac{2\mu^2}{\alpha}(s_r(k-s_c/2) + s_c(k-s_r/2))\right)$$
$$\le \sum_{s_r=1}^{k} \exp\left(s_r\left(\log\frac{k(d-k)e^2}{s_r^2} - \frac{k\mu^2}{\alpha}\right)\right) \sum_{s_c=1}^{k} \exp\left(s_c\left(\log\frac{k(d-k)e^2}{s_c^2} - \frac{k\mu^2}{\alpha}\right)\right).$$

The two terms here are identical, so we will just bound the first one:

$$\sum_{s_r=1}^{k} \exp\left(s_r\left(\log\frac{k(d-k)e^2}{s_r^2} - \frac{k\mu^2}{\alpha}\right)\right)$$

$$\leq \sum_{s_r=1}^{k} \exp\left(s_r\left(\log(k(d-k)e^2) - k\mu^2/\alpha\right)\right)$$

$$\leq k\exp\left(\log(k(d-k)e^2) - k\mu^2/\alpha\right) \qquad \text{if } \mu^2 \geq \frac{\alpha}{k}\log(k(d-k)e^2)$$

Applying this inequality to both terms gives a bound on $W(\mathcal{V}, \alpha)$. This bound is smaller than δ as long as $\mu \geq \sqrt{\frac{c}{k\alpha}\log(k(d-k)e/\delta)}$ for some universal constant c. Again this subsumes the condition required for the inequality to hold.

The other two terms are essentially the same. Using binomial approximations, both expressions can be bounded as:

$$\sum_{s_r=1}^{k} C_k^{s_r} C_{d-k}^{s_r} \exp\left(-\frac{2\mu^2}{\alpha}(s_r k)\right) = \sum_{s_r=1}^{k} \exp(s_r \log(e^2 k(d-k)/s_r^2) - 2s_r k\mu^2/\alpha)$$

$$\leq k\exp(\log(k(d-k)e^2) - 2k\mu^2/\alpha) \qquad \text{if } \mu^2 \geq \frac{\alpha}{2k}\log(k(d-k)e^2).$$

These bounds lead to the same minimax rate as above.

For the lower bound, we again use binomial approximations.

$$W(\mathcal{V}, \alpha) \geq \sum_{s_r=1}^{k}\sum_{s_c=1}^{k} \exp\left(s_r\log\frac{k(d-k)}{s_r^2} + s_c\log\frac{k(d-k)}{s_c^2} - \frac{2\mu^2}{\alpha}(s_r(k-s_c/2) + s_c(k-s_r/2))\right)$$

$$\geq \sum_{s_r=1}^{k}\exp\left(s_r\left(\log\frac{k(d-k)e^2}{s_r^2} - \frac{2k\mu^2}{\alpha}\right)\right)\sum_{s_c=1}^{k}\exp\left(s_c\left(\log\frac{k(d-k)e^2}{s_c^2} - \frac{2k\mu^2}{\alpha}\right)\right)$$

$$\geq \exp(\log(k(d-k) - 2\mu^2 k/\alpha)^2$$

This lower bound goes to infinity if $\mu = o(\sqrt{\frac{1}{k}\log(k(d-k))})$ lower bounds the minimax rate.

To certify that the uniform allocation strategy minimizes $W(\mathcal{V}, \alpha, B)$, we apply Proposition 6. Fix τ and let $\hat{B}$ be such that $\hat{B}((a,b)) = \tau/d^2$ for all $(a,b) \in [d] \times [d]$. By symmetry, every hypothesis achieves the maximum under this allocation strategy, and we will take π to be the uniform distribution over all hypothesis.

For a hypothesis j, let $f_j(B)$ denote the term in the SEDF centered around j. For a hypothesis j based on clusters S_l, S_r and a coordinate (a,b), the subgradient $\frac{\partial f_j(B)}{\partial B(a,b)}$ at $\hat{B}(a,b)$ depends on whether $a \in S_l$ and $b \in S_r$. If $a \notin C_l$ and $b \notin C_r$, then:

$$\left.\frac{\partial f_j(B)}{\partial B(a,b)}\right|_{B=\hat{B}} = \frac{-\mu^2}{\alpha}\sum_{s_r=1}^{k}\sum_{s_c=1}^{k} C_{d-k-1}^{s_r-1} C_k^{s_r} C_{d-k-1}^{s_c-1} C_k^{s_c} \exp(\frac{-2\tau\mu^2}{\alpha d^2}(s_r(k-s_c/2) + s_c(k-s_r/2))).$$

This follows by direct calculation. Similar calculations yield the other cases:

$$\frac{\partial f_j(B)}{\partial B(a,b)}\Big|_{B=\hat{B}} = \frac{-\mu^2}{\alpha} \sum_{s_r=0}^{k-1} \sum_{s_c=1}^{k} C_{d-k}^{s_r} C_{k-1}^{s_r} C_{d-k-1}^{s_c-1} C_k^{s_c} \exp(\frac{-2\tau\mu^2}{\alpha d^2} (s_r(k-s_c) + s_c(k-s_r) + s_r s_c)).$$

$$\frac{\partial f_j(B)}{\partial B(a,b)}\Big|_{B=\hat{B}} = \frac{-\mu^2}{\alpha} \sum_{s_r=1}^{k} \sum_{s_c=0}^{k-1} C_{d-k-1}^{s_r-1} C_k^{s_r} C_{d-k}^{s_c} C_{k-1}^{s_c} \exp(\frac{-2\tau\mu^2}{\alpha d^2} (s_r(k-s_c) + s_c(k-s_r) + s_r s_c)).$$

$$\frac{\partial f_j(B)}{\partial B(a,b)}\Big|_{B=\hat{B}} = \frac{-\mu^2}{\alpha} \sum_{s_r=0}^{k-1} \sum_{s_c=0}^{k-1} C_{d-k}^{s_r} C_{k-1}^{s_r} C_{d-k}^{s_c} C_{k-1}^{s_c} \exp(\frac{-2\tau\mu^2}{\alpha d^2} (s_r(k-s_c) + s_c(k-s_r) + s_r s_c)).$$

These correspond to the cases $a \in S_l, b \notin S_r$, $a \notin S_l, b \in S_r$ and the case where $a \in S_l, b \in S_r$ respectively. The main point is that the value of the subgradient depends only on presence or absence of the row/column in the cluster, and under the uniform distribution π, each row/column is equally likely to be in the cluster. This means that for every coordinate (a,b) taking the expected subgradient with respect to the uniform distribution over hypotheses yields the same expression. So the constant vector is in the subgradient of $f(B)$ at $\hat{B}$, so that $\hat{B}$ is the minimizer of $W(\mathcal{V}, \alpha, B)$ subject to $\|B\|_1 \leq \tau$.

We have already done the requisite calculation to bound the minimax risk under sampling. The calculations above show that if $\mu = \omega(\sqrt{\frac{d^2}{k\tau} \log(k(d-k))})$ then the maximum likelihood estimator, when using the uniform sampling strategy has risk tending to zero. Conversely if $\mu = o(\sqrt{\frac{d^2}{k\tau} \log(k(d-k))})$ then the minimax risk, for *any* allocation strategy tends to one.

The biclusters family is clearly unitarily invariant with respect to the set of orthonormal matrices that permute the rows and columns independently. The family is easiest to describe as acting on the matrices $\mathbf{1}_{S_l} \mathbf{1}_{S_r}^T$. Let P_l, P_r be any two $d \times d$ permutation matrices. Then the matrix $P_l \mathbf{1}_{S_l} (\mathbf{1}_{S_r} P_r)^T$ is clearly another hypothesis, and as we vary P_l and P_r we generate all of the hypothesis. Note that these permutations are unitary operators on the matrix space $\mathbb{R}^{d \times d}$, which allows us to apply Theorem 4.

For the analysis of the interactive algorithm, let us first bound the probability that the algorithm makes a mistake on any single coordinate. Consider sampling a coordinate x with mean μ and noise variance $1/b$. A Gaussian tail bound reveals that:

$$\mathbb{P}[|x - \mu| \geq \epsilon] \leq 2\exp(-2b\epsilon^2).$$

We will sample no more than d^2 coordinates and we will sample each coordinate with the same amount of energy b. So by the union bound, the probability that we make a single mistake in classifying a coordinate that we query is bounded by $\delta/2$ as long as:

$$\mu \geq 2\epsilon = \sqrt{\frac{2}{b} \log(4d^2/\delta)}.$$

We now need to bound b, which depends on the total number of coordinates queried by the algorithm. In the first phase of the algorithm, we sample coordinates uniformly at random until we hit one that is active. Since each sample hits an active coordinate with probability k^2/d^2:

$$\mathbb{P}[\text{hit active coordinate in } T \text{ samples}] = 1 - (1 - k^2/d^2)^T \geq 1 - \frac{1}{e^{Tk^2/d^2}},$$

or if $T = \frac{d^2}{k^2} \log(2/\delta)$, the probability that we hit an active coordinate in T samples will be at least $1 - \delta/2$. The total number of samples we use then can be upper bounded by $2d + \frac{d^2}{k^2} \log(2/\delta)$, which means that we

21

can allocate our budget τ evenly over these coordinates. Therefore we can set $b = \tau(2d + \frac{d^2}{k^2} \log(2/\delta))^{-1}$, and plugging into the condition on μ above proves the result.

Calculations for Hierarchical Clustering: We first describe the heirarchical clustering model.

We study the special case of balanced binary hierarchical clustering on n objects which, without loss of generality we call $[n] = \{1, \dots, n\}$. A binary hierarchical clustering is a collection $\mathcal{C}$ of subset of $[n]$, such $[n] \in \mathcal{C}$, and each $C_\xi \in \mathcal{C}$, if $|C_\xi| > m$, then there exists two sets $C_{\xi \circ L}, C_{\xi \circ R} \in \mathcal{C}$ both of size $C_\xi/2$ that partition C_ξ. As a naming convention, we identify a cluster by a string ξ of L and R symbols. The two sub-clusters of a non-terminal cluster C_ξ are $C_{\xi \circ L}$ and $C_{\xi \circ R}$. The noisy Constant Block Model is defined using this terminology as follows.

Definition 14. *[5] A similarity matrix W is a **noisy constant block matrix** (noisy CBM) if $W \triangleq A + R$ where A is ideal and R is a perturbation matrix:*

- *An **ideal similarity matrix** is characterized by off-block diagonal similarity values $\beta_\xi \in [0,1]$ for each cluster C_ξ such that if $x \in C_{\xi \circ L}$ and $y \in C_{\xi \circ R}$, where $C_{\xi \circ L}$ and $C_{\xi \circ R}$ are two sub-clusters of C_ξ at the next level in a binary hierarchy, then $A_{x,y} = \beta_\xi$. Additionally, $\min\{\beta_{\xi \circ R}, \beta_{\xi \circ L}\} \geq \beta_\xi$. Define $\mu = \min\{\min_\xi\{\min\{\beta_{\xi \circ R}, \beta_{\xi \circ L}\} - \beta_\xi\}, \beta_0\}$, where β_0 is the minimum overall similarity.*

- *A symmetric $(n \times n)$ matrix R is a **perturbation matrix** with parameter σ if (a) $\mathbb{E}(R_{ij}) = 0$, (b) the entries of R are subgaussian, that is $\mathbb{E}(\exp(tR_{ij})) \leq \exp\left(\frac{\sigma^2 t^2}{2}\right)$ and (c) for each row i, $R_{i1}, \dots R_{in}$ are independent.*

We focus on a subfamily of this model, parameterized by n, m, μ where both n and m are powers of two. Our subfamily $\mathcal{H}$ consists of all perfectly balanced hierarchical clusterings on n objects with minimum cluster size m and where all similarities are an integral multiple of μ. This is the simple hierarchical clustering model specified in Section 4. Note that this set of model is a subset of the noisy CBM, so a lower bound for this family applies to the noisy CBM. Let $\mathcal{V}$ denote the class of all such matrices, parameterized by number objects n, minimum cluster size m, and signal strength μ. We interpret $\mathcal{V}$ as a collection of vectors defined by $v_\mathcal{C} = \text{vec}(M[\mathcal{C}])$ for each perfectly balanced hierarchical clustering $\mathcal{C}$.

We now prove Corollary 9. For the first claim, by Theorem 2, we must lower bound the quantity $W(\mathcal{V}, \alpha)$,

$$W(\mathcal{V}, \alpha) = \max_{\mathcal{C} \in \mathcal{H}} \sum_{\mathcal{C}' \neq \mathcal{C}} \exp\left(\|v_\mathcal{C} - v_{\mathcal{C}'}\|_2^2/\alpha\right)$$

Let $\mathcal{C}_0$ be one of these models. Consider perturbing $\mathcal{C}_0$ by taking an object and swapping that object with another one in the adjacent cluster at the deepest level of the hierarchy. There are $nm/2$ such perturbations and any perturbation $\mathcal{C}$ has $\|v_{\mathcal{C}_0} - v_\mathcal{C}\|_2^2 = \mu^2(8m - 4)$. This gives the lower bound of:

$$W(\mathcal{V}, \alpha) \geq \frac{nm}{2} \exp\left(\frac{\mu^2}{\alpha}(8m - 4)\right)$$

By Theorem 2, if $W(\mathcal{V}, 1) \geq 3$, then the minimax risk is bounded from above by $1/2$. Applying our lower bound and solving for μ proves the first part of the result.

For the second claim, if we certify that the uniform sampling strategy minimizes $W(\mathcal{H}, \alpha, B)$ under the budget constraint, then we can immediately apply Theorem 5. We will use Proposition 6 to achieve this.

It is easy to see that for the class $\mathcal{V}$, when B is uniform, every one of these hypotheses achieves the maximum in the definition of $W(\mathcal{V}, \alpha, B)$. Moreover, notice that for every pair of pairs objects $\{a, b\}, \{a', b'\}$,

22

there is a bijection p over $\mathcal{V}$ based on swapping a with a' and b with b' in the hierarchy such that for any hypothesis $v_{\mathcal{C}}$, we have $v_{\mathcal{C}}(a,b) = v_{p(\mathcal{C})}(a',b')$. If in $\mathcal{C}$, a and b are clustered at some level l, then by swapping a with a' and b with b' to form $p(\mathcal{C})$, a' and b' are clustered at level l in $p(\mathcal{C})$ so both terms will be identical because we are in a constant block model.

Since p is a bijection, when we take π to be uniform over the hypotheses, we have:

$$\mathbb{E}_{\mathcal{C} \sim \pi} \sum_{\mathcal{C}' \neq \mathcal{C}} (v_{\mathcal{C}}(a,b) - v_{\mathcal{C}'}(a,b))^2 \exp(-\|v_{\mathcal{C}'} - v_{\mathcal{C}}\|_2^2)$$

$$= \mathbb{E}_{\mathcal{C} \sim \pi} \sum_{\mathcal{C}' \neq \mathcal{C}} (v_{p(\mathcal{C})}(a',b') - v_{p(\mathcal{C}')}(a',b')^2 \exp(-\|v_{p(\mathcal{C}')} - v_{p(\mathcal{C})}\|_2^2)$$

$$= \mathbb{E}_{\mathcal{C} \sim \pi} \sum_{\mathcal{C}' \neq \mathcal{C}} (v_{\mathcal{C}}(a',b') - v_{\mathcal{C}'}(a',b'))^2 \exp(-\|v_{\mathcal{C}'} - v_{\mathcal{C}}\|_2^2).$$

This means we may apply Proposition 6, which certifies that the uniform sampling minimizes the function $W(\mathcal{H}', \alpha, B)$ under budget constraint.

Equipped with this fact, we can reproduce the calculation above but with $B_i = \tau/\binom{n}{2}$, giving:

$$W(\mathcal{V}, \alpha, \tau) \geq \frac{nm}{2} \exp \left(\frac{\mu^2}{\alpha} \frac{\tau}{\binom{n}{2}} (8m - 4) \right)$$

This class is also unitarily invariant using the same tensorized permutation family from the biclustering example. Therefore the MLE is optimal. $\qquad\square$

Calculation for Stars: For the stars problem, define $\mathrm{Nb}(j) \subset V$ to be the neighbors of the vertex j in the graph. For a fixed hypothesis j, we have

$$W_j(\mathcal{V}, \alpha) = \sum_{k \neq j} \exp \left(-\|v_k - v_k\|_2^2/\alpha \right)$$

$$= \sum_{k \in \mathrm{Nb}(j)} \exp(-\mu^2(\deg(k) + \deg(j) - 2)/\alpha) + \sum_{k \notin \mathrm{Nb}(j)} \exp(-\mu^2(\deg(k) + \deg(j))/\alpha)$$

$$\leq \exp \left(-\mu^2 \deg_{\min}/\alpha - \mu^2 \deg(j)/\alpha \right) \left(\deg(j) \exp(2\mu^2/\alpha) + |V| - \deg(j) \right)$$

This last inequality follows by replacing every $\deg(k)$ with $\deg_{\min}$, the lower bound on the degrees. This last expression is maximized with $\deg(j) = \deg_{\min}$, which can be observed by noticing that the derivative with respect to $\deg(j)$ is negative. This gives the bound:

$$W(\mathcal{V}, \alpha) \leq \exp \left(-2\mu^2 \deg_{\min}/\alpha \right) \left(\deg_{\min} \exp(2\mu^2/\alpha) + |V| - \deg_{\min} \right)$$

One can lower bound $W(\mathcal{V}, \alpha)$ by choosing the hypothesis j with $\deg(j) = \deg_{\min}$ and then replacing all other degree terms with $\deg_{\max}$ in the above calculations. This gives:

$$W(\mathcal{V}, \alpha) \geq \exp \left(-\frac{\mu^2}{\alpha} (\deg_{\min} + \deg_{\max}) \right) \left(\deg_{\min} e^{2\mu^2/\alpha} + |V| - \deg_{\min} \right)$$